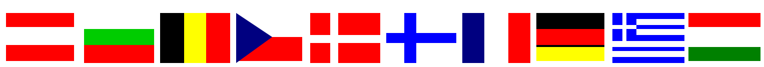
The European Data Grid Project

Ben Segal
CERN
Information Technology Division
B.Segal@cern.ch

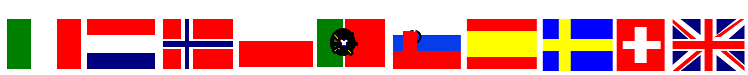
Brian L. Tierney
LBNL



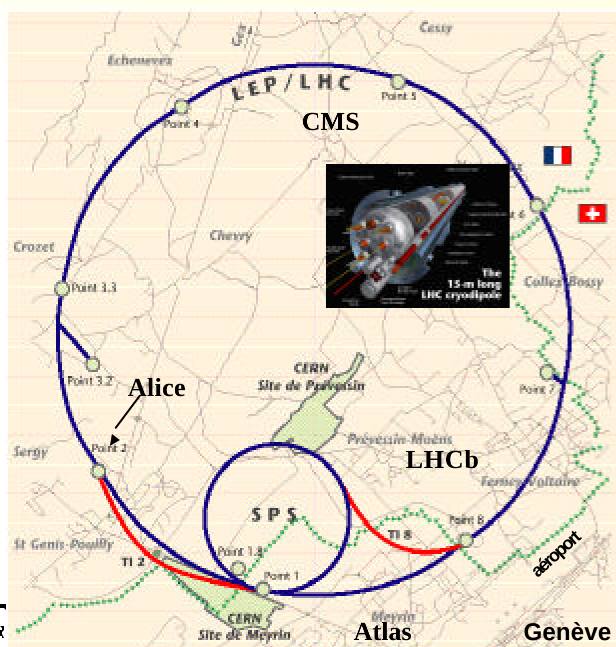
CERN - The European Organisation for Nuclear Research

The European Laboratory for Particle Physics

- Fundamental research in particle physics
- Designs, builds & operates large accelerators
- Financed by 20 European countries
- SFR 950M budget - operation + new accelerators
- 3,000 staff
- 6,000 users (researchers) from all over the world
- Experiments conducted by a small number of large collaborations:
  LEP experiment (finished) : 500 physicists, 50 universities,
  20 countries, apparatus cost SFR 100M
  LHC experiment (future ~2005) : 2000 physicists, 150 universities, apparatus costing SFR 500M



CERN



The LHC at CERN is an accelerator which brings protons and ions into head-on collisions at higher energies than ever achieved before. This will allow scientists to penetrate still further into the structure of matter and recreate the conditions prevailing in the early universe, just after the "Big Bang"

The LEP accelerator

- World's largest particle collider, ran for 11 years.
- 27 km circumference, 100 m underground
- Counter circulating beams of electron positron bunches
- Four experiments have confirmed Standard Model predictions to high precision
- Maximum collision energy of 209 GeV

Many questions still remain ? LHC

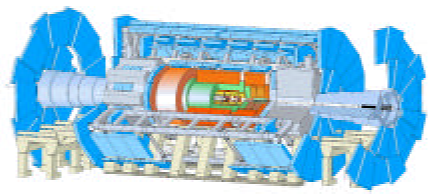
LHC in the LEP Tunnel

- Counter circulating beams of protons in the same beampipe.
- Centre of mass collision energy of 14 TeV.
- 1000 superconducting bending magnets, each 13 metres long, field 8.4 Tesla.
- Super-fluid Helium cooled to 1.9° K

World's largest superconducting structure

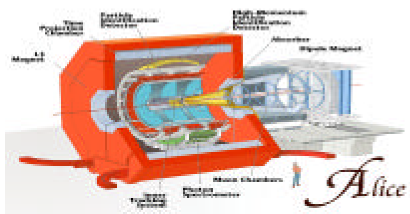
The LHC detectors

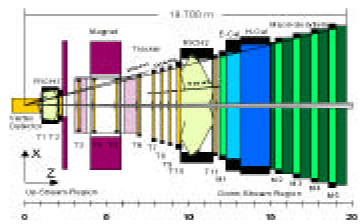
CMS



ATLAS

3.5 Petabytes/year ? 10^9 events/year



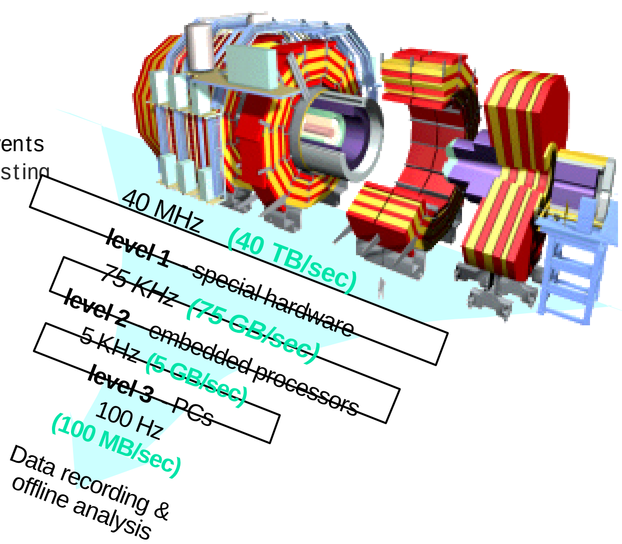


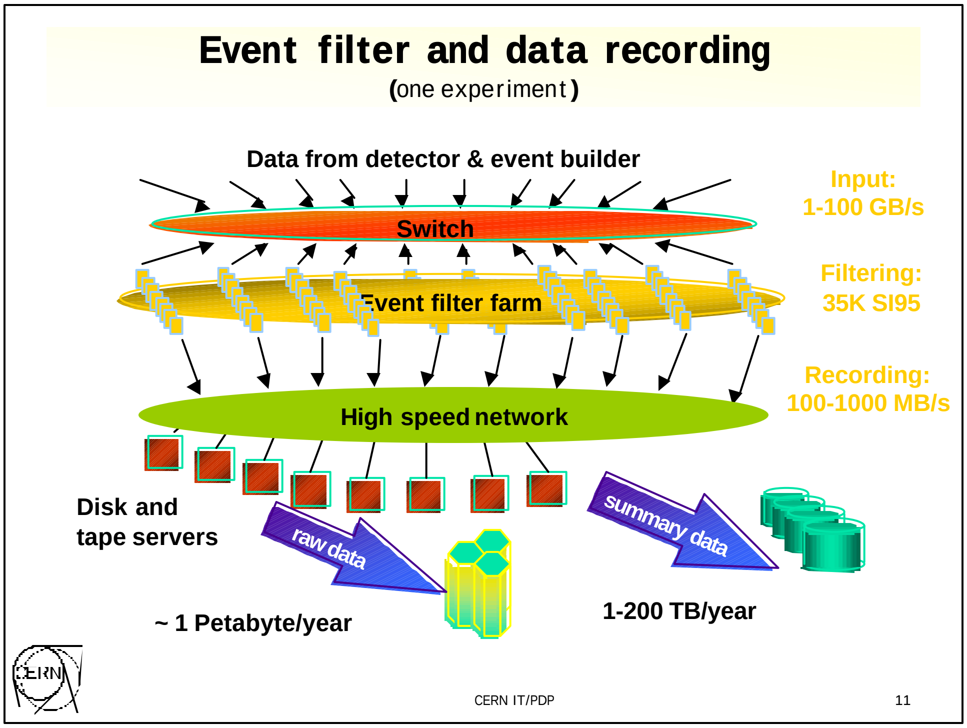
LHCb

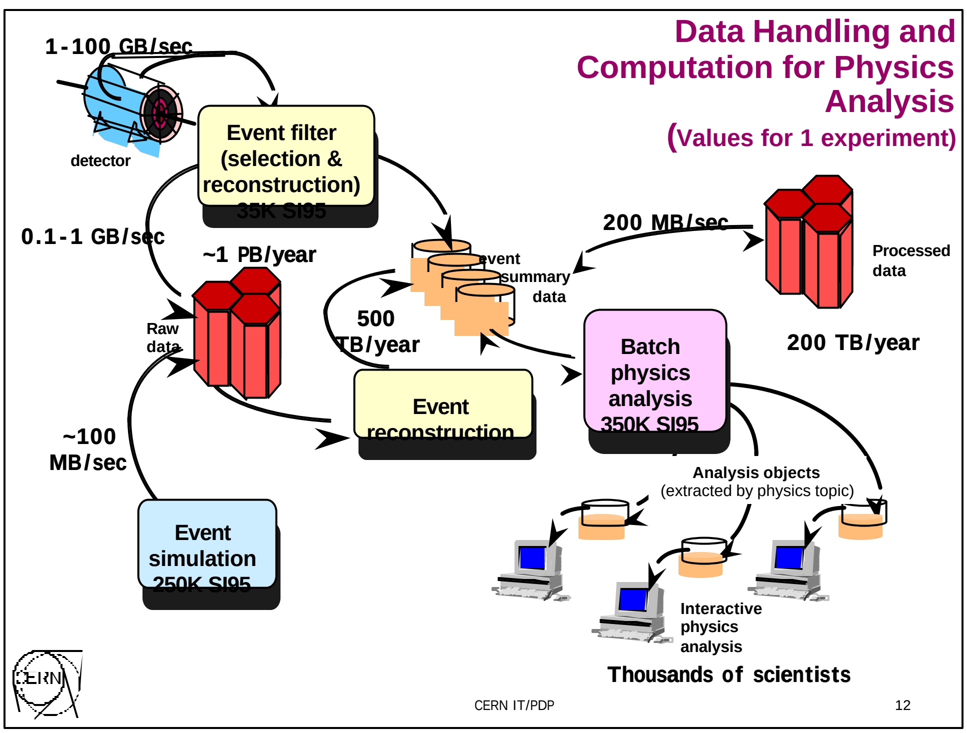
Online system

C M S
Compact Muon Solenoid

- Multi-level trigger
  - Filter out background
  - Reduce data volume
  - Online reduction 10^7
- Trigger menus
  - Select interesting events
  - Filter out less interesting




Event filter and data recording
(one experiment)
Data from detector & event builder
Input:
1-100 GB/s
Switch
Filtering:
35K SI95
Event filter farm
Recording:
100-1000 MB/s
High speed network
Disk and
tape servers
raw data
summary data
~ 1 Petabyte/year
1-200 TB/year
CERN IT/PDP
11


Data Handling and
Computation for Physics
Analysis
(Values for 1 experiment)
1-100 GB/sec
detector
Event filter
(selection &
reconstruction)
35K SI95
0.1-1 GB/sec
~1 PB/year
200 MB/sec
Processed
data
event
summary
data
Raw
data
500
TB/year
Batch
physics
analysis
350K SI95
200 TB/year
Event
reconstruction
~100
MB/sec
Analysis objects
(extracted by physics topic)
Event
simulation
250K SI95
Interactive
physics
analysis
Thousands of scientists
CERN IT/PDP
12

LEP to LHC

Parameter	LEP	COMPASS	LHC
Raw data rate	1MB/sec	35MB/sec	100MB/sec
Number of events/year	$<10^7$	$\sim10^{10}$	$\sim10^9$
Raw data volume/year	0.2-0.3 TB	300TB	1 PB
Event size	20 – 50 kB	30kB	1 MB
Event reconstruction time	2–8 Si95-secs	2 Si95-secs	500 Si95-secs
Number of users	400 - 600	~200	~2000
Number of institutes	30-50	~35	~150

Each LHC experiment requires one to two orders of magnitude greater than the TOTAL capacity installed at CERN today

All LEP: < 1TB/year Rate: 4MB/sec
All LHC: ~ 3PB/year Alice rate: 1GB/sec

Characteristics of HEP computing

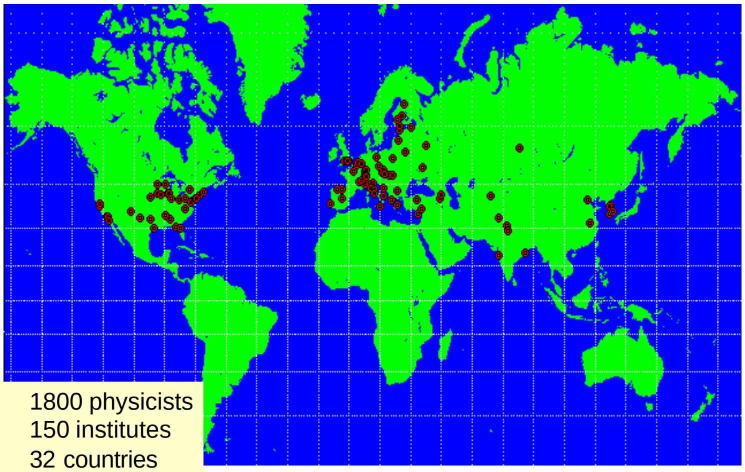
Event independence
- Data from each collision is processed independently
- Mass of independent problems with no information exchange

Massive data storage
- Modest event size: 1-10 MB
- Total is very large - Petabytes for each experiment.

Mostly read only
- Data never changed after recording to tertiary storage
- But is read often ! cf.. magnetic tape as an archive medium

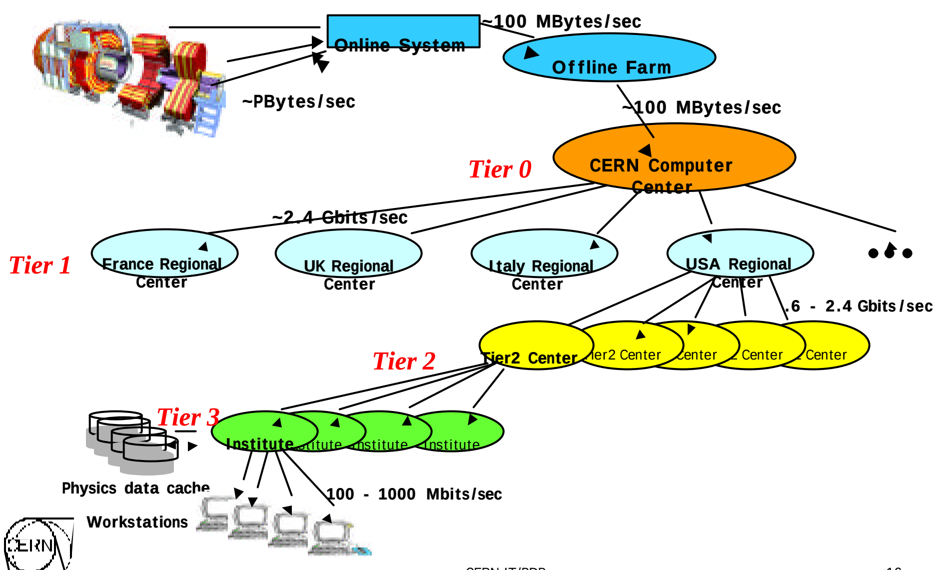
Modest floating point needs
- HEP computations involve decision making rather than calculation
- Computational requirements in SPECint95 secs

World Wide Collaboration
? distributed computing & storage capacity



CMS: 1800 physicists
150 institutes
32 countries

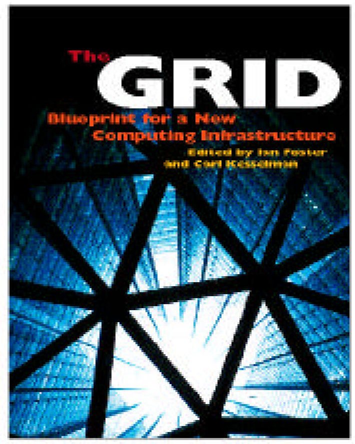
LHC Multi-Tier Model



Can Grid technology be applied to LHC computing?

The GRID metaphor

- Analogous with the electrical power grid
- Unlimited ubiquitous distributed computing
- Transparent access to multi peta byte distributed data bases
- Easy to plug in
- Hidden complexity of the infrastructure



Ian Foster and Carl Kesselman, editors, "The Grid: Blueprint for a New Computing Infrastructure," Morgan Kaufmann, 1999, http://www.mkp.com/grids

GRID from a services view

Applications

- Chemistry
- Cosmology
- Environment
- Biology
- High Energy Physics

Application Toolkits

- Distributed computing toolkit
- Data-intensive applications toolkit
- Collaborative applications toolkit
- Remote Visualisation applications toolkit
- Problem solving applications toolkit
- Remote instrumentation applications toolkit

Grid Services (Middleware)

Resource-independent and application-independent services
authentication, authorisation, resource location, resource allocation, events, accounting, remote data access, information, policy, fault detection

Grid Fabric (Resources)

Resource-specific implementations of basic services
E.g., transport protocols, name servers, differentiated services, CPU schedulers, public key infrastructure, site accounting, directory service, OS bypass

What does the Grid do for you?

- You submit your work
- And the Grid
  - Finds convenient places for it to be run
  - Organises efficient access to your data
    - Caching, migration, replication
  - Deals with authentication to the different sites that you will be using
  - Interfaces to local site resource allocation mechanisms, policies
  - Runs your jobs
  - Monitors progress
  - Recovers from problems
  - Tells you when your work is complete
- If there is scope for parallelism, it can also decompose your work into convenient execution units based on the available resources, data distribution

European Data Grid partners

Managing partners

UK: PPARC Italy: INFN France: CNRS Netherlands: NIKHEF
ESA/ESRIN CERN

Industry

IBM (UK), Compagnie des Signaux (F), Datamat (I)

Associate partners

Istituto Trentino di Cultura (I), Helsinki Institute of Physics / CSC Ltd (FI), Swedish Science Research Council (S), Zuse Institut Berlin (DE), University of Heidelberg (DE), CEA/DAPNIA (F), IFAE Barcelona, CNR (I), CESNET (CZ), KNMI (NL), SARA (NL), SZTAKI (HU)

Other sciences

KNMI(NL), Biology, Medicine

Formal collaboration with USA being established

Preliminary programme of work

Middleware

WP 1 Grid Workload Management	INFN
WP 2 Grid Data Management	B. Segal/CERN
WP 3 Grid Monitoring services	R. Middleton/PPARC
WP 4 Fabric Management	O. Barring/CERN
WP 5 Mass Storage Management	J. Gordon/PPARC

Grid Fabric -- testbed

WP 6 Integration Testbed	F. Etienne/CNRS
WP 7 Network Services	C. Michau/CNRS

Scientific applications

WP 8 HEP Applications	F. Carminati/CERN
WP 9 EO Science Applications	L. Fusco/ESA
WP 10 Biology Applications	C. Michau/CNRS

Management

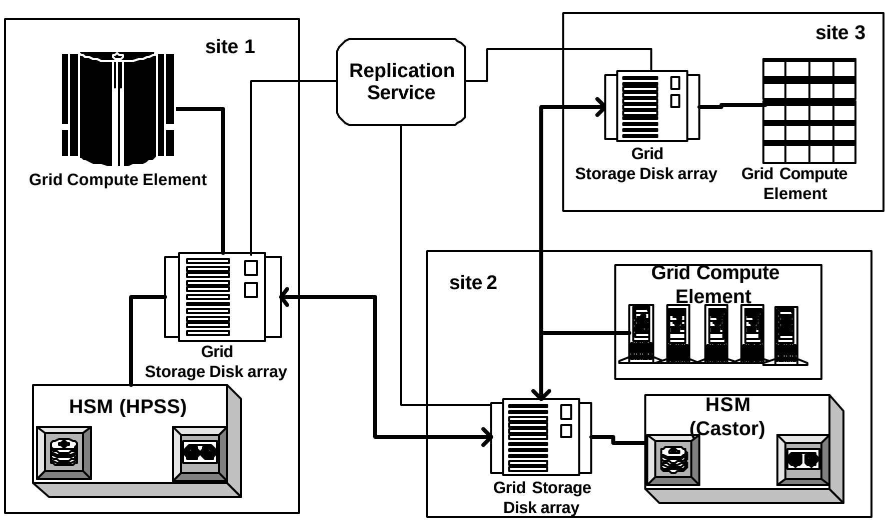
WP 11 Dissemination	G. Mascari/CNR
WP 12 Project Management	F. Gagliardi/CERN

EU Data Grid Main Issues

- Project is by EU standards very large in funding and participants
- Management and coordination is a major challenge
- Coordination between national (European) and EU Data Grid programmes
- Coordination with US Grid activity (GriPhyN, PPDG, Globus)
- Coordination of the HEP and other sciences objectives
- Very high expectations already raised, could bring disappointments

25

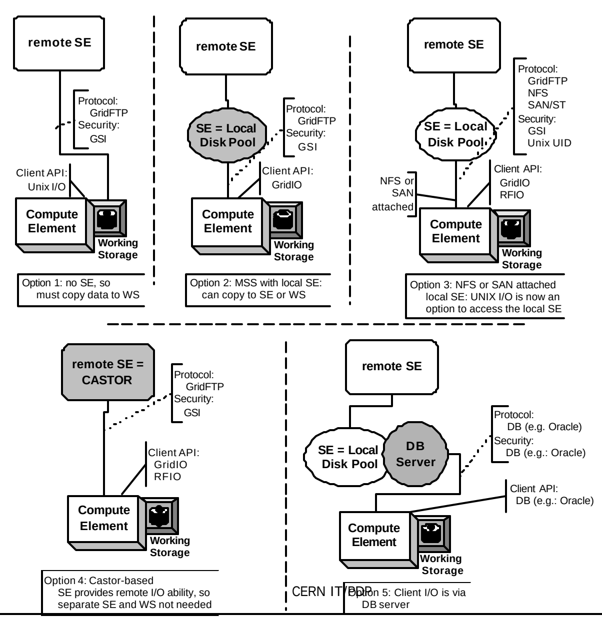
Data Grid Storage Model



Terminology

- StorageElement:
  - **any storage system with a Grid Interface**
    - **supports GridFTP**
- Logical File Name (LFN)
  - **globally unique**
  - **LFN://hostname/string**
    - **hostname = virtual organization id**
      - **use of hostname guarantees uniqueness**
    - **e.g.: LFN://cms.org/analysis/run10/event24.dat**
- Physical File Name (PFN)
  - **PFN://hostname/path**
    - **hostname = StorageElement host**

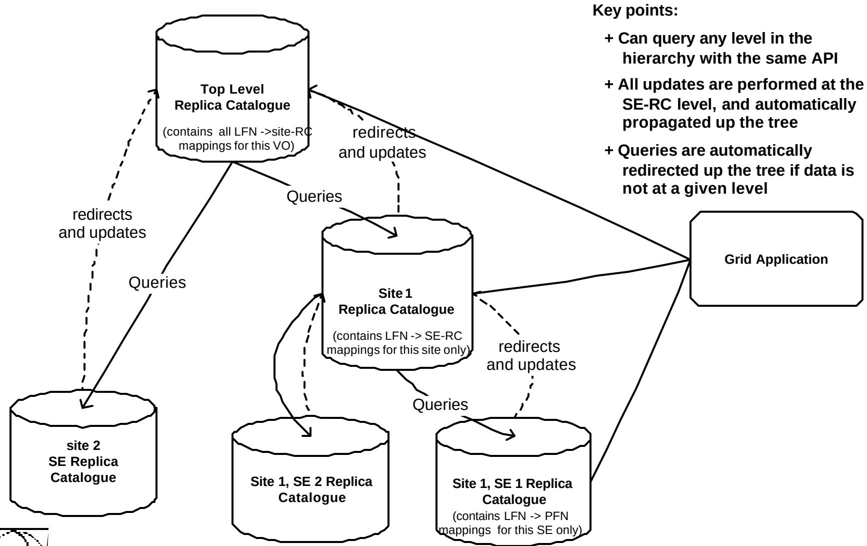
SE to CE Issues



WP2 work: Replica Catalog

- Replica Catalog contains a table mapping LFN to 1 or more PFNs.
- There should be exactly one ReplicaCatalogue for each StorageElement
- All Replica Catalogues for a given virtual organization are linked together, probably in a tree structure
  - **"leaf" catalogues contain a mapping of LFN to PFN**
  - **"non-leaf" catalogues contain only a pointer to another replica catalogue**
- All ReplicaCatalogues (leaf and non-leaf) have identical client APIs

Sample Hierarchy



Key points:

+ Can query any level in the hierarchy with the same API
+ All updates are performed at the SE-RC level, and automatically propagated up the tree
+ Queries are automatically redirected up the tree if data is not at a given level

WP2 Work: Replica Management

- Reliable File Copy
- Cost Estimation
- StorageElement Control
  - **allocation**
  - **pinning**

End